IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Personalized Recipe Recommendation System using Hybrid Approach

Lipi Shah[1], Hetal Gaudani[2], Prem Balani[3]

PG Student, Dept of Information Technology, G H Patel College of Engineering and Technology, Vidhyanagar, India [1]

Associate Professor, Dept of Computer Engg, G H Patel College of Engineering and Technology, Vidhyanagar, India [2]

Assistant Professor, Department of Information Technology, G H Patel College of Engineering and Technology, Vidhyanagar, India [3]

Abstract: This paper describes the recommendation system in culinary domain. With the prevalence of internet, whole world is connected and different users of different countries are sharing millions of new recipes on the internet, world widely. So, as a result users are not aware of the all the recipes on the web. Recipe contains different heterogeneous information's like ingredients, cooking procedure, categories etc. So, we think the recipe is aggregation of the different heterogeneous features. Most of the recommendation system is based on the content or collaborative filtering to predict the new recipe of interest for a user. Incorporating with the both the filtering techniques, we present an effective and elegant framework for combining both techniques in recipe recommendation system. Most of the recipe recommendation system uses content information as ingredients or cooking procedures of recipes. We proposed the hybrid approaches by incorporating conventional techniques, content as well as collaborative filtering, by adding more heterogeneous information of recipes like cuisines, preparation direction, dietary etc. and try to reduce RMSE than the conventional recommendation system.

Keywords: Recommendation system, collaborative filtering, hybrid approaches, recipes, content information.

I. INTRODUCTION

This is the era of the internet, so enormous amount of data are deposited on the web every day. On the web, large number of choices is available, so information filtering is required to extract useful information from these raw data available on the web. In the culinary domain, currently 10,000 websites [1] are available on the web provide the different types of information for filtering purpose like photos, texts, videos etc. But, large amount of information are deposited on the web by different users of different countries make information overwhelming. So, finding useful recipe which may be liked by the user will be time consuming process. So, here recipe recommendation system provides the desirable solution for that problem.

Many on-line stores are provided recommendation system like Amazon, CD NOW and Netflix etc. Most of they are using the prevalence approaches that is content based and collaborative filtering techniques. These all websites are much popular on the web so, the review for the product on the website is higher than the recipes website, which is less popular. So, the task of recipe recommendation system in culinary domain is providing the some of the challenges. First are the Sparsity problem, that is the recipe website is not much popular as other websites so, the rating for the particular recipe is very much less than that popular websites. Second is the recipe is aggregations of the large number of features like ingredients, cooking directions, cooking methods, categorical information. So, here large number of heterogeneous data is available. To, understand any recipe, these heterogeneous information is needed. For understand complex opinion of the recipes with the user's perspective, all these heterogeneous information is needed. Most of the conventional recommendation system is heavily depends on the collaborative filtering to find the latent connection among different users of the same website using user's rating matrix. But there are some problems like cold start, Sparsity, popularity bias, first rater proem etc. So, overcome those problem, we propose the new hybrid approaches which used both content as well as collaborative filtering informations in the personalized recipe recommendation system. To navigate the users on the web based on their past preferences, we propose two hybrid approaches using contents of recipe and user's rating matrix for recipes. The first hybrid approach is based on the KNN collaborative filtering technique with the recipe content information. The second hybrid approach is based on the stochastic gradient descent approach with user's rating information of recipes as well as content information of recipes.

In this paper, section II. Describes the existing system, section III. Describes the flow of the system and statistics of dataset, section IV. Describes the proposed hybrid approaches, section V. describes the evaluation result and section VI. Describes the conclusion and future work.

II. EXISTING SYSTEM

The root of the recommendation system is the result of the extensive work in cognitive science [4], approximation theory [9], information retrieval [5], forecasting theories

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

[14], and recommendation system emerged as an independent area in the mid-1990s. There are many real world applications available that use the recommendation system to help their users to find more appropriate products or items at users point of view and increase the production as well as profit at business point of view. In the culinary domain, many researches are tackled in recommendation system of recipes in the past. Yoko et al. [6] propose a recipe recommendation system which is based on user's schedule, weight of user etc. Farhana et al. [16] propose the personalized cancer diet panning based on the different nutrition requirement to the cancer patients and provide tips chart. Peter Forbes et al. [7] propose a recipe recommendation system based on the different ingredients and finding similar ingredients for making different constitutions. Tsuguya et al. [10] propose the recipe recommendation system based on the natural language process (NLP) using nutrition information of the recipe. Dalwinderjeet Kaur et al. [8] propose the network of ingredients of different cultures; co-relate the different ingredients of different cultures. Liping Wang et al. [19] propose the substructure similarity of different preparation direction of recipes and finding similar types of recipes. Jill Freyne et al. [13] propose the recipe recommendation system based on the different ingredients of the recipes. Takuma Maruyama et al. [17] propose a recipe recommendation system based on the recipe ingredients object reorganization and suggest different recipes based on the ingredients reorganization. Yong-Yeol Ahn et al. [11] propose the flavor network for food-pairing and make substitute ingredients. Mayumi Ueda et al. [18] propose a recipe recommendation system based on the ingredients and the quantity of ingredients in the recipes. In the all past researches, the recipes are aggregation of ingredients and nutrition value are considered only. In our propose work, we have added many heterogeneous information of ingredients in our models like preparation directions, occasions, cuisines, dietary etc. We are taking conventional models as benchmark models and compare our propose models with them.

III.FLOW OF THE SYSTEM AND STATISTICS OF DATASET

This section is describing the flow of the general recommendation system of recipe recommendation system and information is extracted from the dataset.

A. Flow Of the System

The general flow of our system is shown in figure1. The web crawler program is needed to extract the information of user's rating and the recipes information from the web. Here the information is taken from the food.com website. After the crawling process, the pre-processing step is taken place and stored the useful data into the recipe content database. After crawling, for each user, now we have the sparse rating matrix that is of user versus recipe. After that, divide the sparse user versus recipe matrix into train and test dataset. After that, train the recommendation using training dataset. Apply the test dataset on the train model and evaluation and prediction is done for the model. The dataset is extracted from 25-2-2000 to 9-3-2012 food.com website. From this data, we get different recipes and their content information as well as user's rating for the recipes in the range of 1 to 5.

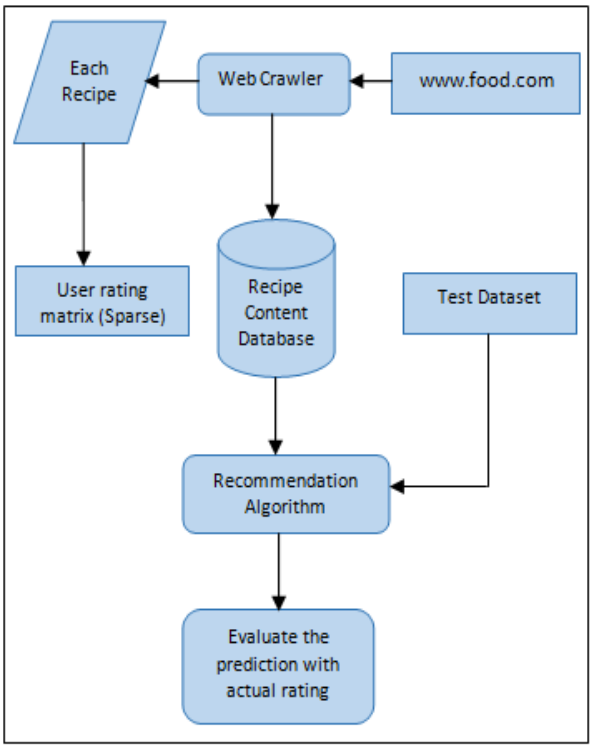

Fig. 1 General flow of the system

In the content informations of the recipes, we get different categorical data like preparation direction, dietary information, different cuisines, occasional information and different courses information. These all informations are used as content information of the each recipe. From the dataset, different statistic data and the features information are obtained which is described section III-B.

B. Statistics Of Dataset

After the crawling process, the data is pre-process. In which, we remove the recipes which only contain three or less time ratings. Using the string matching function, similar types of ingredients are made into the single ingredient. So, here many-to-one relationship is taken consider. For example, Cortland apple, dried apple, green apples converted to apples. After pre-processing step, the statistical data is obtained which is shown in the table I, table II and table III.

TABLE I. STATISTICS OF DATASET

Number of Recipes	10,971
Number of Users	23,807
Total available ratings	3,43,308
Sparsity	0.132%
Average rating/ user	14.42
Average rating/ item	31.29

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

TABLE II. STATISTICS OF INGREDIENTS

Total ings. counts in all recipes	56,740
Max. ings. in a recipe	40
Min. ings. in a recipe	2
Average ings. in a recipe	6
Max. appearance of ings. in a recipe	2282
Min. appearance of ings. in a recipe	1
Average appearance of ings. in a recipe	13.05

TABLE III STATISTICS OF FEATURES EXCLUDING INGREDIENTS

Total feats. counts in all recipes	2,41,259
Max. feats. in a recipe	88
Min. feats. in a recipe	3
Average feats. in a recipe	21.68
Max. appearance of feats. in a recipe	10,909
Min. appearance of feats. in a recipe	1
Average appearance of feats. in a recipe	106.75

Here, the all statistics information of the dataset is given. In the system total 23,807 users and 10,971 recipes are available. Total available ratings are 3, 43,308. So, from that the Sparsity of the system can be calculated and that is 0.132%. In the other two tables, the features like ingredients, preparation directions, cuisines, courses etc. statistics are given.

IV. PROPOSED HYBRID APPROACHES

The aim of this work is to find which algorithm is suitable for personalized recipe recommendation by comparing the evaluation parameter RMSE of different conventional algorithms. We have focused on two types of data. In these approaches, we divided each recipe into different features like ingredients, different preparation directions like less than 30 minutes, three steps or less, less than 60 minutes etc., different cuisines like Indian, Asian, American, Italian etc., different dietary like low fat, high protein, high carbohydrate etc., different occasions like diwali, birthday, summer, dinner party etc., different courses like dessert, main dishes, salad etc...One is fine grained data of recipes for their different features and second is the high level data that is rating for recipes by each user. Here we have proposed, two hybrid approaches using content and rating informations on recipes which is described in section 4-A and 4-B.

A. Proposed Hybrid Approach-1

This hybrid approach is based on the KNN (k-nearest neighbor) algorithm with use of the content informations of recipes and rating information of users on recipes. So, here one fine grain information of content of recipe is used which is implicit information gathering process and second rating of users on recipe which is explicit information gathering process. So, two matrixes are generated.

Using fine grained data, content matrix C is generated:

C = recipeid × featureid / ingredientid

$$C = \begin{cases} 1, & \forall\ ings, features\ \epsilon\ recipeid_{train} \\ 0, & Otherwise \end{cases}$$

Using rating of users on the recipes, user's rating matrix R is generated:

R = userid × recipeid

Using these two matrices C and R the algorithm is worked as follow, which in shown in flow diagram in Figure: 2.

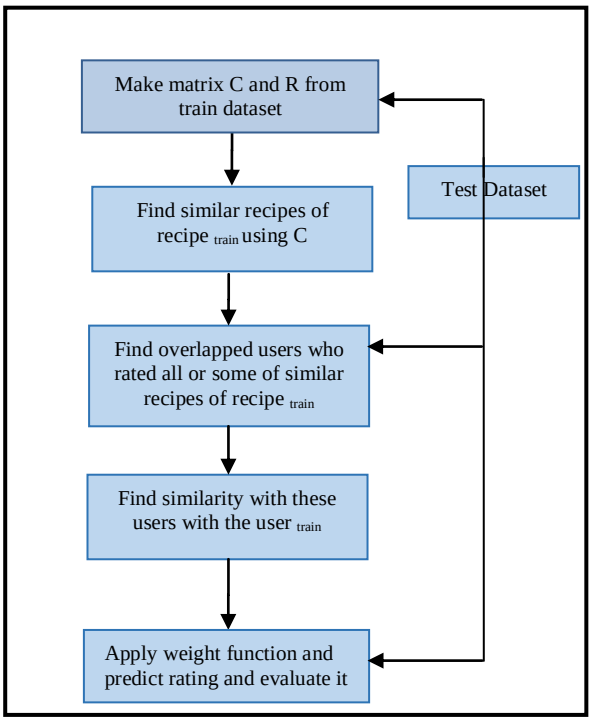

Fig. 2 Flow of Proposed Hybrid approach-1

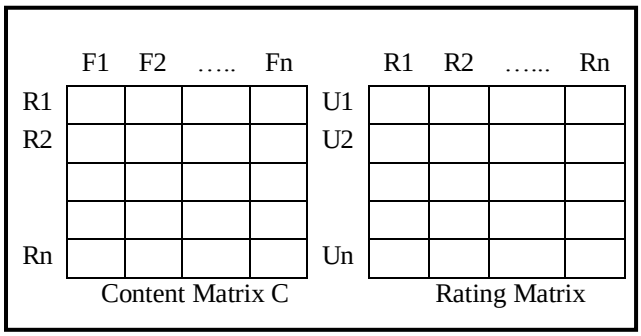

FIG. 3 Structure of content Matrix C and rating Matrix R

For given $recipeid_{train}$, find the similar type of other recipes from the content matrix C.

Similarity between recipes using KNN approach using Content matrix C:

$$\text{sim}(r_a, r_b) = \frac{\sum_{i=1}^{N} (r_{a_i} - \overline{r_a})(r_{b_i} - \overline{r_b})}{\sum_{i=1}^{N} (r_{a_i} - \overline{r_a})^2 \sum_{i=1}^{N} (r_{b_i} - \overline{r_b})^2} \quad (1)$$

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Here r_a and r_b indicate the recipe a and recipe b respectively, r_{a_i} indicates the recipe a contains features f (including ingredient I) or not in 0 or 1 binary form. $\overline{r_a}$ indicates the average of recipe a (average of entire raw of recipe a) in content matrix C. Here N indicates the total number of recipes in the system.
Similarity between users using KNN approach using user's rating matrix R:

$$\text{sim}(u_a,u_b) = \frac{\sum_{i=1}^{N} (u_{a_i} - \overline{u_a})(u_{b_i} - \overline{u_b})}{\sum_{i=1}^{N} (u_{a_i} - \overline{u_a})^2 \sum_{i=1}^{N} (u_{b_i} - \overline{u_b})^2} \quad (2)$$

Here u_a and u_b indicate the user a and user b respectively, u_{a_i} indicates the user a contains recipe i or not in the range of 1 to 5. Here, $\overline{u_a}$ indicates the average of user a (average of entire raw of user a) in rating matrix R. Here N indicates the total number of users in the system.
The weight function for finding rating of recipe for given user based on the similarity thresholding value:

$$\text{rat}(u_a, r_i) = \frac{\sum_{n \in Thresholdinglimit} sim(u_a,u_n)\, rat(u_n,r_i)}{\sum_{n \in Thresholdinglimit} sim(u_a,u_n)} \quad (3)$$

Here, $\text{rat}(u_a, r_i)$ indicates the rating prediction for user u_a for recipe id i that is r_i, based on the different thresholding values between 0.1 to 0.9. Here u_n indicates the users who satisfied the particular thresholding values.

B. Proposed Hybrid Approach-2
This hybrid approach is based on the SGD (Stochastic gradient descent) algorithm with the content information as well as rating information of users on the recipes. SGD is the one of the collaborative filtering method and model based method. So, here the model is created first and based on that the prediction is done. Here only one matrix is used which contain fine grained content information of recipe and user's rating of recipes. The matrix is defined as:

Make matrix A, which contains users and recipe contain information.

$$A = \begin{cases} User\ train\ rating\ table\ (recipeid, userid) = rating \\ 1, \quad \forall\ ings, features\ \epsilon\ recipeid_{train} \\ 0, \quad Otherwise \end{cases}$$

Using matrix A, the flow of the algorithm is shown in figure 4. The matrix A defined by both information rating as well as content. The structural view of matrix is shown in figure 5.

For given $recipeid_{train}$ and $userid_{train}$, first the matrix A is created with rating and content information of recipes R. After that SGD (stochastic gradient descent) approach is applied here on rating matrix A is explained here. Using SGD, matrix A is divided into two matrices p and q. p and q is two sub-matrixes, whose multiplicative score try to give original matrix A which try to add some predicted values on the empty cell of matrix A.

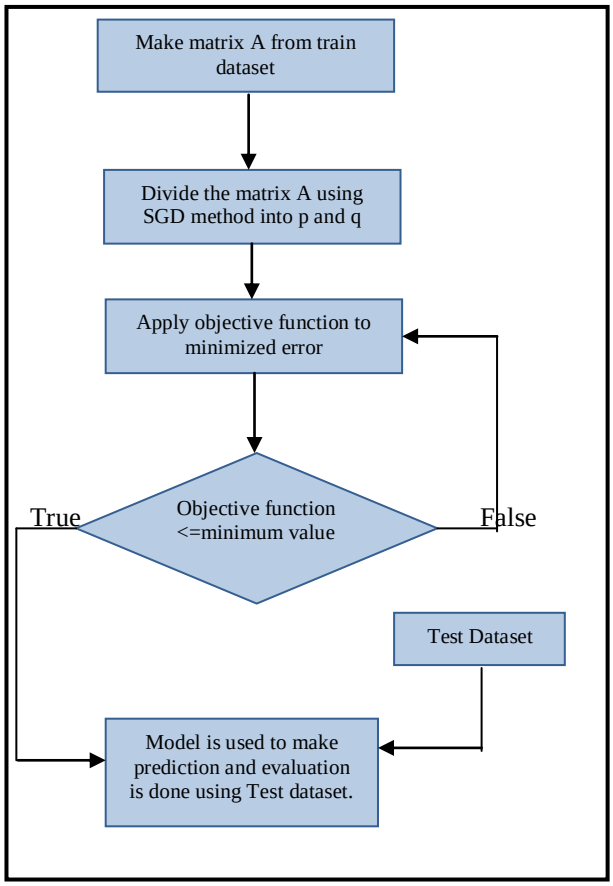

Fig. 4 Flow of Proposed Hybrid approach-2

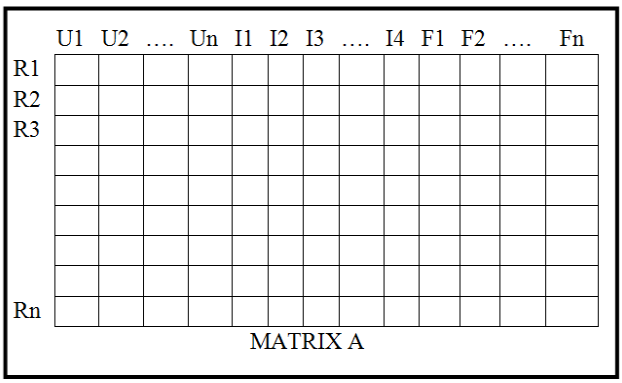

Fig. 5 Structure of matrix A

$$\bar{r}_{ui} = q_i^T p_u \quad (4)$$

The objective (loss) function is defined below is used to minimized the value of two sub-matrices p and q to reduced the error.
Apply the value of the q and p into the objective function:

$$\min_{q^*,p^*} \sum_{(u,i)\in k} (r_{u,i} - q_i^T p_u)^2 + \lambda (||q_i||^2 + ||p_u||^2) \quad (5)$$

λ is the regularization parameter, k is the training samples, $r_{u,i}$ is rating of the training samples, p_u is user hidden factors and q_i is recipes hidden factors.
This objective function is checked every time with predefined minimum value and model is created and

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

evaluation is done. To find the p and q value is iterative process which is used the objective function. So the value of p and q is based on the previous value of p and q which is defined as:

$$q_i' \leftarrow q_i + \gamma . (e_{ui} . p_u - \lambda . q_i) \quad (6)$$
$$p_u' \leftarrow p_u + \gamma . (e_{ui} . q_i - \lambda . p_u) \quad (7)$$

Here γ is the step size, e_{ui} is the error, q_i' is new value based on the q_i and p_u' is new value based on the p_u.

V. EVALUATION

There are different approaches in RS available previously from 1990. From that, here we have taken three conventional models as our benchmarks that are: User-user based collaborative filtering, Item-item based collaborative filtering and rating based SGD (Stochastic gradient descent). We have compared our proposed hybrid approaches which contain the rating informations as well as content information with the conventional three models in culinary domain.

A. Setup

There are about 10,000 websites are available for recipe to extract each recipe's information with rating provided by users. For that we have created the crawler program for that which is extracting JSON data from web and we have extracted our required information from that. Crawling process is long time processes to extract such a large data. We required about 10-15 days for extracting these web contents. From this information, we got user's rating on different recipes, each recipe's content informations like different ingredients and different features. As a pre-processing after cleaning the data, we have found features/ingredients co-occurrences to find the importance of ingredients/ features in the system. After that, we have taken most ninety co-occurrences of ingredients as well as most ninety co-occurrences of features of recipes as our content information. So, we finally got the content information and rating information of recipes in the system, which we used in our proposed models.

B. Results

To find the performance of the models, RMSE (Root mean square error) is used here rather than MAE (Mean absolute error). RMSE is very common and it makes an excellent general purpose error metric for numerical predictions. Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors. Here we have taken, three conventional models as our benchmarks, comparison is shown in figure: 6 with RMSE of the all five models.

From the figure: 6, the first model that is user-user collaborative filtering gives 0.7703 RMSE. This model only uses the user's rating information on recipes and finds the prediction of rating based on the users' similarity. he second model, item-item based collaborative filtering T reduces the RMSE to 0.6848 compared to user-user based collaborative filtering.

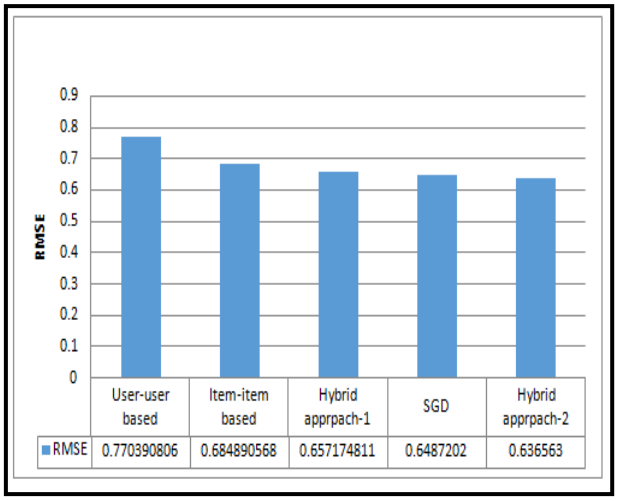

Fig. 6 Comparison of models with RMSE parameter

This model also uses only user's rating information based on the items similarity. In our proposed hybrid model-1, we have used content information with rating information and reduce the RMSE to 0.6571. Here the model works based on the users similarity plus content similarity. As the SGD model is one of the powerful two fold model as it first create the pre-computed model and find the rating prediction after that. In the fourth model, SGD reduces the RMSE to 0.6487 which only uses the rating information of the users on recipes. In the fifth model, proposed Hybrid approach-2 reduces the RMSE to 0.6365 which uses the content information of recipes as well as user's rating information on recipes.

VI. CONCLUSION AND FUTURE WORK

Recommendation system is the progressive field in the last decade, when numbers of content-based, collaborative and hybrid approaches are proposed for different industrial growth purposes. In our work, we choose the culinary domain here. Both the content as well as collaborative filtering has their advantages as well as disadvantages. So to overcome each other's problems, here we proposed hybrid approaches and try to reduce RMSE. SGD based hybrid approach gives better RMSE than others. Here the features and the ingredients of recipes are broken down and used as content information. 1% of reduction of RMSE is also gives lots of effectiveness into RS.

Here we used around ten thousand of recipes and twenty three thousand of user, as a future work, the large dataset can be used as well as more features can be added like the flavours of ingredients by chemical properties for deep filtering purpose so that different tastes and substitute ingredients can be defined more finer.

REFERENCES

1. http://www.alexa.com/topsites/category/Top/Home
2. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, (8), 30-37.
3. Wang, L., Li, Q., Li, N., Dong, G., & Yang, Y. (2008, April). Substructure similarity measurement in chinese recipes.

In Proceedings of the 17th international conference on World Wide Web (pp. 979-988). ACM.

4. E. Rich, "User Modeling via Stereotypes," Cognitive Science, vol. 3, no. 4, pp. 329-354, 1979.
5. G. Salton, Automatic Text Processing. Addison-Wesley, 1989.
6. Mino, Y., & Kobayashi, I. (2009, November). Recipe recommendation for a diet considering a user's schedule and the balance of nourishment. InIntelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on (Vol. 3, pp. 383-387). IEEE.
7. Forbes, P., & Zhu, M. (2011, October). Content-boosted matrix factorization for recommender systems: experiments with recipe recommendation. InProceedings of the fifth ACM conference on Recommender systems (pp. 261-264). ACM.
8. Kular, Dalwinderjeet Kaur, Ronaldo Menezes, and Eraldo Ribeiro. "Using network analysis to understand the relation between cuisine and culture."Network Science Workshop (NSW), 2011 IEEE. IEEE, 2011.
9. M.J.D. Powell, Approximation Theory and Methods. Cambridge Univ. Press, 1981.
10. Ueta, Tsuguya, Masashi Iwakami, and Takayuki Ito. "Implementation of a goal-oriented recipe recommendation system providing nutrition information."Technologies and Applications of Artificial Intelligence (TAAI), 2011 International Conference on. IEEE, 2011.
11. Ahn, Yong-Yeol, et al. "Flavor network and the principles of food pairing."Scientific reports 1 (2011).
12. Zhang, R., Liu, Q. D., Gui, C., Wei, J. X., & Ma, H. (2014, November). Collaborative Filtering for Recommender Systems. In Advanced Cloud and Big Data (CBD), 2014 Second International Conference on (pp. 301-308). IEEE.
13. Freyne, Jill, and Shlomo Berkovsky. "Intelligent food planning: personalized recipe recommendation." Proceedings of the 15th international conference on Intelligent user interfaces. ACM, 2010.
14. J.S. Armstrong, Principles of Forecasting—A Handbook for Researchers and Partitioners. Kluwer Academic, 2001.
15. Linden, Greg, Brent Smith, and Jeremy York. "Amazon. com recommendations: Item-to-item collaborative filtering." Internet Computing, IEEE 7.1 (2003): 76-80.
16. Saludin, F. A., Zakaria, N., & Husain, W. (2010, November). User requirement analysis for personalized cancer dietary planning and menu construction. In Biomedical Engineering and Sciences (IECBES), 2010 IEEE EMBS Conference on (pp. 410-416). IEEE.
17. Maruyama, T., Kawano, Y., & Yanai, K. (2012, November). Real-time mobile recipe recommendation system using food ingredient recognition. InProceedings of the 2nd ACM international workshop on Interactive multimedia on mobile and portable devices (pp. 27-34). ACM.
18. Ueda, M., Asanuma, S., Miyawaki, Y., & Nakajima, S. (2014). Recipe recommendation method by considering the user's preference and ingredient quantity of target recipe. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1).
19. Wang, Liping, et al. "Substructure similarity measurement in chinese recipes." Proceedings of the 17th international conference on World Wide Web. ACM, 2008.
20. De Campos, Luis M., et al. "Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks."International Journal of Approximate Reasoning 51.7 (2010): 785-799.